Overlapping Waves in Tool Use Development: a Curiosity-Driven Computational Model

Sébastien Forestier, Pierre-Yves Oudeyer

Overlapping Waves in Tool Use Development: a Curiosity-Driven Computational Model

Sébastien Forestier
Université de Bordeaux and Inria Bordeaux Sud-Ouest
Email: sebastien.forestier@inria.fr

Pierre-Yves Oudeyer
Inria Bordeaux Sud-Ouest and Ensta Paristech
Email: pierre-yves.oudeyer@inria.fr

***Abstract*—The development of tool use in children is a key question for the understanding of the development of human cognition. Several authors have studied it to investigate how children explore, evaluate and select alternative strategies for solving problems. In particular, Siegler has used this domain to develop the overlapping waves theory that characterizes how infants continue to explore alternative strategies to solve a particular problem, even when one is currently better than others. In computational models of strategy selection for the problem of integer addition, Shrager and Siegler proposed a mechanism that maintains the concurrent exploration of alternative strategies with use frequencies that are proportional to their performance for solving a particular problem. This mechanism was also used by Chen and Siegler to interpret an experiment with 1.5- and 2.5-year-olds that had to retrieve an out-of-reach toy, and where they could use one of several available strategies that included leaning forward to grasp a toy with the hand or using a tool to retrieve the toy. In this paper, we use this domain of tool use discovery to consider other mechanisms of strategy selection and evaluation. In particular, we present models of curiosity-driven exploration, where strategies are explored according to the learning progress/information gain they provide (as opposed to their current efficiency to actually solve the problem). In these models, we define a curiosity-driven agent learning a hierarchy of different sensorimotor models in a simple 2D setup with a robotic arm, a stick and a toy. In a first phase, the agent learns from scratch how to use its robotic arm to control the tool and to catch the toy, and in a second phase with the same learning mechanisms, the agent has to solve three problems where the toy can only be reached with the tool. We show that agents choosing strategies based on a learning progress measure also display overlapping waves of behavior compatible with the one observed in infants, and we suggest that curiosity-driven exploration could be at play in Chen and Siegler's experiment, and more generally in tool use discovery.**

I. Introduction

The development of tool use in children is one of the essential questions for the understanding of the global development of human cognition. Indeed, children progressively learn to interact with multiple objects in varied ways which shows an understanding of objects' shapes, relations, the action of forces, and other physical properties used for mental transformations and planning: crucial tools to human cognition.

Different theories have been developed to explain child development, beginning with a description of successive stages in development by Piaget [1]. Piaget described the developmental stages as necessary behaviors that infants should display in a given order. Related views proposed that in the context of reasoning, a single child reasons with only one method at any given age. However, different views were developed more recently which describe more variability in the possible developmental paths that children are driving. Siegler's overlapping waves theory [2] is describing and modeling the way infants represent and select a set of currently available methods to solve a problem. In the overlapping waves theory, infants maintain their set of methods (also called strategies) with associated frequencies depending on the past history of the use of those methods. The frequencies evolve over time while new strategies are discovered which explain the observed changes in behavior. For instance, when learning the mathematical addition, infants use different methods from one trial to another, and may continue to use non-optimal methods for a long period of time even if they already know more efficient methods. Siegler has speculated that such continued exploration of alternative and sub-optimal methods to solve a family of problem may be useful to acquire skills that will later facilitate the resolution of new problems. This cognitive variability could be an essential mechanism to acquire greater knowledge, which might be more important for learning in childhood than just having high quality performances on specific tasks.

Siegler et al developed several computational models of strategy selection and evolution to account for how children learn how to add integer numbers: ASCM (Adaptive Strategy Choice Model [2]), and SCADS (Strategy Choices and Strategy Discoveries [3]). Those models are argued to closely parallel the development of addition strategies with the use of several strategies, with errors in the execution of those strategies. In SCADS, furthermore, a mechanism allows the discovery of new strategies and the authors show that the same strategies are discovered and in the same sequences as with children. In the two models, the strategies are selected with frequencies that are directly proportional to (called a "matching law" on) their success rate in the corresponding previous problems. This model also included a novelty bias to explore new strategies more than their success rate would allow: the value for exploring new strategies was initialized optimistically (then decreasing in time if success rate did not rise). The focus of this model has been the mode of strategy selection (matching law), with a measure of the value of strategies based on their performance to solve a given task.

However, these models have not considered other forms of value systems, such as curiosity-driven information seeking, which could play a key role in child development [4].

In the context of tool use development, Chen and Siegler [5] conducted an experiment with 1.5- and 2.5-year-olds that had to retrieve an out-of-reach toy with one of the six available tools. Children were exposed to a sequence of three similar problems with different tool shapes and visual features, but for each problem only one tool was effective to retrieve the toy. They designed three conditions. In the control condition, the mother just asked the child to get the toy. In the hint condition, the experimenter moreover suggested to use the target tool. Finally, in the modeling condition, the experimenter actively showed to the infant how to retrieve the toy with the target tool. First, they show that in the control condition only few children succeeded to retrieve the toy with the tool even after three problems (less than 10% of the 1.5-year-olds and less than 20% of the 2.5-year-olds). However, in the hint condition and modeling conditions, a large proportion of 1.5-year-olds and most of the 2.5-year-olds succeeded to use the tool strategy by the end of the experiment. With respect to the strategic variability, the authors measured that 74% of toddlers used at least three strategies. The different strategies observed were to lean forward and try to retrieve the toy with the hand (forward strategy), to grab one of the tool and try to catch the toy with the tool (tool strategy), to ask to the mother if she could retrieve the toy for them (but she was told not to) or to walk around the table to look at the toy from different angles (indirect strategy), and finally some of the children did not engage in any of those strategies (no strategy).

Firstly, the authors reported the dynamics of strategy choice as an average over children. They showed that the tool strategy frequency was on average increasing with the successive trials and the forward strategy was decreasing in the hint and modeling conditions, whereas in the control condition the tool strategy remained stable. This pattern was interpreted by the authors as a clear overlapping waves pattern besides the fact that it was a pattern of the average over children. The overlapping waves theory suggests that this pattern of strategy change should be visible on a per child basis, meaning that each child should always use a set of strategies and smoothly change their frequency use. However, the observed average pattern does not imply that each child (or most of them) display an overlapping waves pattern. It could be that in Chen and Siegler's experiment, each child begins with the forward strategy, and at some point in the experiment (different for each child), switch to the tool strategy and never uses again the forward one. In that case, an average of the strategy use would also show a smooth increase in the tool strategy and decrease in the forward strategy use. Nevertheless, the authors also reported a measure that could disentangle the different hypothesis [5, p42]. They measured the average proportion of trials where children used other strategies than the tool strategy after the first trial where they used the tool strategy. The toddlers in the control condition did use the other approaches than the tool strategy on more than half the trials after the first time they used the tool strategy (84% of the trials for 1.5-year-olds, 48% for 2.5-year-olds). In contrast, in the hint and modeling conditions, the young infants used other approaches in around 20% of the trials, and older infants in only 4%. This result showed that strategic variability did continue after children began to use the tool strategy in the control condition but not in the hint and modeling conditions. Consequently, we do not agree with the conclusions of the authors saying that a clear overlapping waves pattern was visible regarding the change in forward versus tool strategy use. According to this analysis, overlapping behaviors were observed in this experiment only in the control condition where the mother just asked the infant to retrieve the toy, and the experimenter did not add further incentive.

In this paper, we consider the problem of the modeling of overlapping waves of behaviors in the context of tool learning and use. We will target to model alternative mechanisms that could be at play in Chen and Siegler's experiment. In addition this model will also be used to model learning and strategy selection mechanisms happening before the experiment (hence modeling learning of tool use taking place "at home" during the months preceding the lab sessions). These unified learning mechanisms will be used for both free play exploration/learning of tool use (from scratch) and for exposure to evaluation in lab sessions with an incentive to solve a task. Indeed, a source of difficulty to interpret the results of behavioral experiments in babies is that it is difficult to control for what happened before the lab sessions. In particular, we can't know exactly how much prior experience the toddlers had playing with objects and tools at home, what kind of tools were available, and how the caregivers were interacting with the child or answering its requests to get toys. Furthermore, understanding how the object saliency and the cues of the caretaker are interpreted by the children is an open question. The interpretation of these experiments has implicitly assumed that the experimental setup was designed so that the children would "want" to catch the toy (this also applies to similar experiments such as [6]). However, as we will suggest through the model below, alternative hypotheses can be considered (and be non-exclusive). In particular, we will suggest that a salient object may trigger curiosity-driven exploration, where the child explores to gain information about which strategy allows to get it (rather than trying to maximize its probability to actually catch it).

We build upon a previous model of curiosity-driven development of tool use in a simulated 2D environment with objects and tools [7]. Intrinsic motivations, or "curiosity", have been shown to play a fundamental role in driving spontaneous exploration in infant free play [9]. Intrinsic motivations have been defined as mechanisms that push infants to explore activities for their own sake, driven by the search of novelty, surprise, dissonances or optimal challenge [10]. This model is learning different sensorimotor models structured in a hierarchy that represents the environmental structure. The use of an intrinsic motivation for the exploration of sensorimotor mappings yielding a high learning progress allowed the emer-

gence of a smooth progression between overlapping phases of behavior similar to the one found in infants [8]. In the model, the intrinsic motivation self-organized a first phase where the agents were mainly exploring movements of the arm without touching objects, then the exploration of the interaction with a single object, and finally a smooth shift towards behaviors experimenting the interaction of multiple objects.

In the present paper, we use a similar model and study different mechanisms for adaptively selecting alternative strategies to reach a toy, which were not studied in our previous work focused on evaluating the impact of hierarchical representations of sensorimotor spaces [7]. We hypothesize that not only do the type of decision mechanism to select an action (matching law, or greedy: choose the best one) can influence the resulting behavior and match observations in infants as explained in Siegler's models, but also the measure on which the decision is based, whether it is a competence measure, as in ASCM and SCADS, or an information-gain based measure such as learning progress.

To test this hypothesis, we designed an experimental setup with two phases. In the first one, the agents are autonomously exploring their environment through three sensory spaces (representing the hand, stick and toy), and can learn how to move their hand, how to grab an available stick, and how to reach a toy with either the hand or the stick. In a second phase, the agents use the same strategy selection procedure as in the first phase, but are now only exploring towards retrieving the toy, which mimics the incentive given by the mother to retrieve the toy in Siegler's lab experiment [5]. In Siegler's experiment, several tools where available but only one allowed to grab the toy, and the tool strategy was defined as trying to use any of the tool to reach for the toy. We simplify this setup and we place only one tool in the environment so that the tool strategy only contains one type of actions and is easier to interpret. We measure the success rates to grab the toy and we study the evolution of the use of tool and hand strategies in this second phase depending on the mechanism of strategy selection, for individual agents.

Together with this paper, we provide open-source Python code[1] with Jupyter notebooks explaining how to reproduce the experiments and analysis.

II. Methods

A. Environment

We simulate a 2D robotic arm that can grasp a block or grasp a stick and use it to move the block. In each trial, the agent executes a motor trajectory and gets the associated sensory feedback. At the end of each trial, the arm and the stick are reset to their initial state, and the block is reset to a random location every 20 iterations. The next sections precisely describe the items of the environment and their interactions. See Fig.1 for an example state of the environment.

[1]Source code and Jupyter notebooks available as a Github repository at https://github.com/sebastien-forestier/ICDL2016

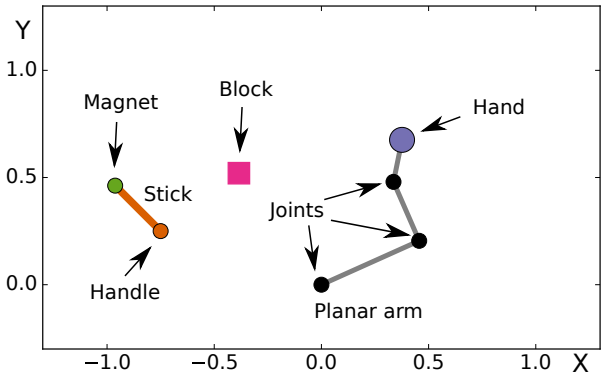


Fig. 1. A state of the environment. The initial position of the arm is vertical so in this position the first and third joints are rotated to the right and the second joint to the left. The magnetic stick is at its initial position and is reset at each iteration. The block can be caught either by the magnetic side of the stick or directly by the hand as it is reachable here. The block is only reset every 20 iterations to a random position reachable by the hand.

1) Robotic Arm: The 2D robotic arm has 3 joints. Each joint can rotate from $-\pi$ to π (rad) around its resting position, which is seen by the agent as a standard interval $[-1, 1]$. The length of the 3 segments of the arm are 0.5, 0.3 and 0.2 so the length of the arm is 1 unit. The resting position of the arm is vertical with all joints at 0 rad and its base is fixed at position $(0, 0)$. A trajectory of the arm will be represented as a sequence of vectors in $[-1, 1]^3$.

2) Motor Control: We use Dynamical Movement Primitives [11] to control the arm's movement as this framework permits the production of a diversity of arm's trajectories with few parameters. Each of the 3 arm's degrees-of-freedom (DOF) is controlled by a DMP starting at the rest position of the joint. Each DMP is parameterized by one weight on each of 2 basis functions and one weight specifying the end position of the movement. The weights are bounded in the interval $[-1, 1]$ and allow each joint to fairly cover the interval $[-1, 1]$ during the movement. Each DMP outputs a series of 50 positions that represents a sampling of the trajectory of one joint during the movement. The arm's movement is thus parameterized with 9 weights, represented by the motor space $M = [-1, 1]^9$.

3) Objects: A stick and a toy (block) are available in the environment. The stick can be grasped by the handle side and can be used as a tool to catch the block. The stick has length 0.3 and is initially located at position $(-0.75, 0.25)$ as in Fig. 1. If the hand reaches the block (within 0.2), we consider that the block is grasped until the end of this movement. Similarly, if the hand reaches the handle side of the stick (within 0.1), the stick is considered grasped and follows the hand's position with the direction of the arm's last segment until the end of this movement. If the magnetic side of the stick reaches the block (within 0.1), then the block follows the stick's magnet.

4) Sensory Feedback: At the beginning of each trial, the agent gets the context of the environment: the position of the block ($Context$, 2D). At the end of the movement, it gets sensory feedback from the following items in the environment.

First, the trajectory of the hand is represented as its x and y positions at 3 time points: steps 17, 33, 50 during the movement of 50 steps (S_{Hand}, 6D). Similarly, the trajectory of the magnet of the stick is a 3-point sequence of x and y positions (S_{Stick}, 6D). It also gets the initial and final position of the block, and the minimal distance during the movement between the hand and the block, if the stick was not grasped, or between the magnet and the block, if the stick was grasped (S_{Block}, 5D). The total sensory space S has 17 dimensions.

B. Learning Agent

The problem settings for the learning agent is to explore its sensorimotor space and collect data so as to generate a diversity of effects in the three available sensory spaces, and to learn inverse models to be able to reproduce those effects. In this section we describe the hierarchical learning architecture.

1) Global Architecture of Sensorimotor Models: The agent learns 4 sensorimotor models at the same time (see Fig. 2). Model 1 learns a mapping from the motor space M to S_{Hand}, model 2 from S_{Hand} to S_{Stick}, model 3 from S_{Hand} to S_{Block} and model 4 from S_{Stick} to S_{Block}. The block is the only item that can have a different initial position at the beginning of each iteration. We thus call contextual models the two models that have to take into account this context (models 3 and 4), and non-contextual models the two others (models 1 and 2). Those two types of models provide the inverse inference of a probable motor command m (in their motor space) to reach a given sensory goal s (in their sensory space), but their implementation is slightly different (see next sections).

In order to get interesting data to build its sensorimotor model, the agent performs Goal Babbling. It first chooses one of the three sensory spaces, and then self-generates a goal in the sensory space and tries to reach it. To generate those goals, different strategies have been studied [12]. Here we use a random generation of goals for the exploration of spaces S_{Hand} and S_{Stick} (Random Goal Babbling), which was proven to be highly efficient in complex sensorimotor spaces [13]. For S_{Block}, we just define the goal as moving the block to the origin position $(0, 0)$.

If the goal is in S_{Block}, the agent also has to decide which method to use in order to try to retrieve the block: either the forward method, with model 3, or the tool method with model 4. In the other cases, if the goal is chosen in S_{Hand} or S_{Stick}, then model 2 or respectively 3 is used. Once the babbling model is chosen, it performs inverse inference and uses lower-level models to decide which motor command m will be experimented in the environment.

Finally, when motor parameters m have been tested in the environment and feedback s received, the mappings of models 1 and 2 are updated, and if the agent grasped the tool, then model 4 is updated, otherwise model 3 is updated. Also, a measure of success to reach the goal and of learning progress are computed and will be used to help choosing the space to explore. We use the Explauto autonomous exploration library [14] to define those sensorimotor models and the learning progress measure.

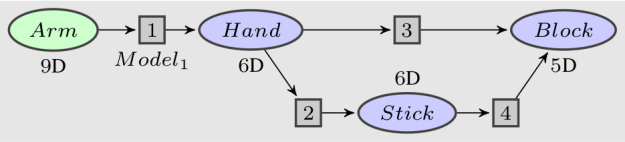


Fig. 2. Architecture of models. The green circle is the motor space and the blue ones are sensory spaces. The gray squares are the 4 models.

2) Non-Contextual Models: Each non-contextual model has a motor space (e.g. motor space of model 2 is S_{Hand}) and a sensory space (respectively S_{Stick}). They learn a mapping and provide the inverse inference of a probable motor command m (in its motor space) to reach a given sensory goal s (in its sensory space). They store new information of the form (m, s) with $m \in M$ being the experimented motor parameters and $s \in S_i$ the associated sensory feedback in their sensory space. They compute the inverse inference with the nearest neighbor algorithm: they look at the nearest neighbor in the database of a given s in the sensory space, and return its associated motor parameters. Model 1 also adds exploration noise (gaussian with $\sigma = 0.01$) to explore new motor parameters.

3) Contextual Models: The inverse inference is computed differently for contextual models (models 3 and 4). Whatever the position of the block (context), the agent tries to grasp it (with the hand for model 3 and with the tool for model 4) and to put it at the origin location $(0, 0)$. To do so, if the context is new (not within 0.05 of any previously seen context), then the agent chooses the motor command that in the past led to the grasping of the block in the closest context. If the context is not new, then the model chooses the sensory point in the database with the smallest cost among the points that had a similar context (context within 0.05 of the current one), and a gaussian noise ($\sigma = 0.01$) is added to the motor position. The cost of a sensory point s_{block} with context c is

$$cost(c, s_{block}) = D_{S_b}(traj, c) + D_{S_b}(origin, p_{final}) \quad (1)$$

where $D_{S_{block}}(traj, c)$ was the minimal distance between the hand (for model 3) or tool (model 4) and the toy during the trajectory. Also, $origin$ is the position $(0, 0)$ and p_{final} is the final position of the toy. Finally, D_{S_i} is a normalized distance in a sensory space S_i,

$$D_{S_i}(s, s') = \frac{||s - s'||}{max_{s_1,s_2}||s_1 - s_2||} \quad (2)$$

4) Active Space Babbling: At each iteration, the architecture first has to choose the sensory space S_i to explore. This choice is probabilistic and proportional to the interest of each space (but with $\epsilon = 5\%$ of random choice). We call this procedure Active Space Babbling.

When space S_{Hand} is chosen to be explored, a random goal s_g (hand trajectory) is sampled and then sensorimotor model 1 is used to infer a motor command m to realize this hand trajectory. When space S_{Stick} is chosen, a random goal s_g (stick trajectory) is sampled and model 2 is used to infer a

hand trajectory to make this stick trajectory (and model 1 used to realize the hand trajectory). When space S_{Block} is explored, then model 3 or 4 (hand or tool strategy) has to be chosen (see next section) to reach for the toy and the goal s_g is to catch the toy and put it at position $(0, 0)$.

We now define the learning progress and interest of a sensorimotor model mod that tries to reach the goal s_g (e.g. model 1 if S_{Hand} was chosen, or model 4 if S_{Block} and the stick were chosen). Once the motor command m is executed, the agent observes the current sensory feedback s in the chosen sensory space S_i. This outcome s might be very different from s_g as this goal can be unreachable, or because lower-level models are not mature enough for that goal. We define the progress $P(s_g)$ associated to the goal $s_g \in S_i$:

$$P(s_g) = D_{S_i}(s_g, s') - D_{S_i}(s_g, s) \tag{3}$$

where s_g and s are the current goal and reached sensory points, and s'_g and s' are the previous goal of the model mod that is the closest to s_g, and its associated reached sensory point. The progress of model mod is initialized at 0 and updated to follow the progress of its goals (with rate $n = 1000$):

$$P_{mod}(t) = \frac{n-1}{n}\ P_{mod}(t-1) + \frac{1}{n}\ P(s_g) \tag{4}$$

where t is the current iteration. The interest of model mod is its absolute progress, meaning that a negative progress is also interesting:

$$I_{mod}(t) = |P_{mod}(t)| \tag{5}$$

Now we define the interest of space S_{Hand} and S_{Stick} as the interest of models 1 and 2 respectively. The interest of space S_{Block} is the sum of the interest of models 3 and 4.

5) Choice of Method to Reach the Block: When the agent has chosen to explore S_{Block}, and given a block position (context), it has to choose one of its two available methods to reach the block: the hand method (model 3) or the tool method (model 4). We define 4 conditions with different choices, based on two measures; competence and interest. The competence measure estimates for each method if the agent will be able to grasp the block. It is computed as follows: if the block was never grasped with the method, then it is -1, otherwise it is the distance of the closest context where the block was grasped. The interest measure estimates the learning progress of each method to reach the current block position. If the context is strictly new, then the interest is the inverse distance of the closest context where the block was grasped (or 1 if there was no such context). If the context is not new, which means that the block was not grasped in the previous attempts, then the interest is computed as a derivative of the costs of the previous attempts for this context. If there were n previous attempts a_i, then the interest is

$$\left|\ mean_{\frac{n}{2}+1..n}\ [cost(a_i)] - mean_{1..\frac{n}{2}}\ [cost(a_i)]\ \right| \tag{6}$$

where the cost of an attempt is the one of Section II-B3. Finally, for each of those two measures, we define two types of choice for both measures. The ϵ-greedy choice is a random choice with probability $\epsilon = 5\%$, and the choice of the highest with probability $(1 - \epsilon)$. In the matching law choice, the probability of choosing each method is proportional to the measure, but also with $\epsilon = 5\%$ probability of a random choice. This results in 4 possible conditions:

- GC: greedy on competence
- MC: matching law on competence
- GI: greedy on interest
- MI: matching law on interest

C. Experiments

The experimental procedure is composed of two phases. In phase 1, the agents are autonomously learning for 1000, 2000, 5000 or 10000 iterations where we reset the toy to a random position (but reachable directly with the hand) each 20 iterations. In phase 2, the agents are successively exposed to 3 new problems (or contexts) while they keep updating their sensorimotor models, for 200 iterations allowed per new problem. In those 3 problems, the toy is set at a location reachable with the tool but not reachable with the hand (problem A: $(-0.1, 1.2)$, B:$(0, 1.25)$, C:$(0.1, 1.2)$). Those locations are distinct enough so that given the solution to one of them, solving another one requires some exploration, but close enough so that the previous one helps. Finally, we simulate 100 independent trials for each condition.

III. Results

Fig. 4 shows an example of the evolution of the interests to explore the sensory spaces during phase 1 for an agent of condition MI. After some iterations, the interest of S_{Block} becomes greater than the interest of S_{Hand} and S_{Stick} and thus is more often chosen to be explored. Fig. 3 shows in each condition and for each of the 3 problems of phase 2, the proportion of the 100 agents that succeeded to reach the toy, depending on experience (the number of iterations performed in phase 1, i.e. the number of sensorimotor experiments/movements already achieved by each agent). We see that in all conditions and for all problems, the success rate increases with experience. For instance, for problem A in condition MI, the success rate goes from 25% when agents have experimented 1000 iterations to 50% when they have experimented 10000. Also, for all conditions and experiences, the success rate increases from problem A to B and from problem B to C. For example, the success rate is 21% for problem A of condition MC at experience 1000 and it goes to 27% for problem B and 33% for problem C. Finally, the success rates of all problems in condition GC are smaller by 5 to 20% than the success rates of the three other conditions, and the success rates of condition MI are slightly higher than those of condition MC.

Fig. 5 shows 2D maps of the preference between the hand and tool strategies to reach the block depending on its 2D position (on a 100×100 grid), for one agent of experience 10000 iterations of each condition that succeeded to catch the block on the three problems. Also, the maps are computed at different times of phase 2 for each condition: at the beginning of phase 2 before problem A, after problem A, after problem

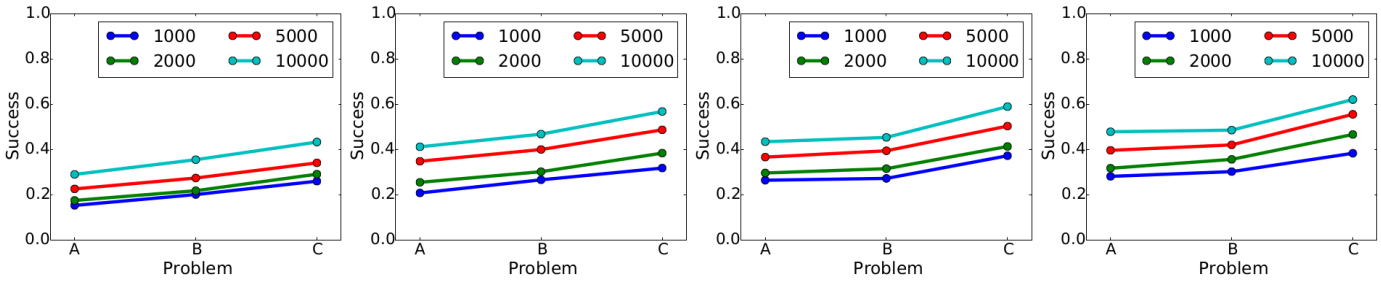


Fig. 3. Proportion of 100 agents that succeeded to reach the toy in each of the 3 problems of phase 2, depending on condition and experience (the number of iterations experimented). Success rate increases with experience and with the problems encountered, and are better in conditions MC, GI and MI than GC.

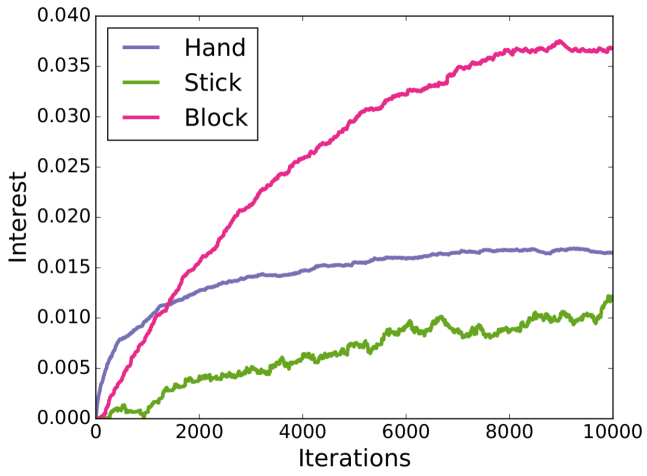


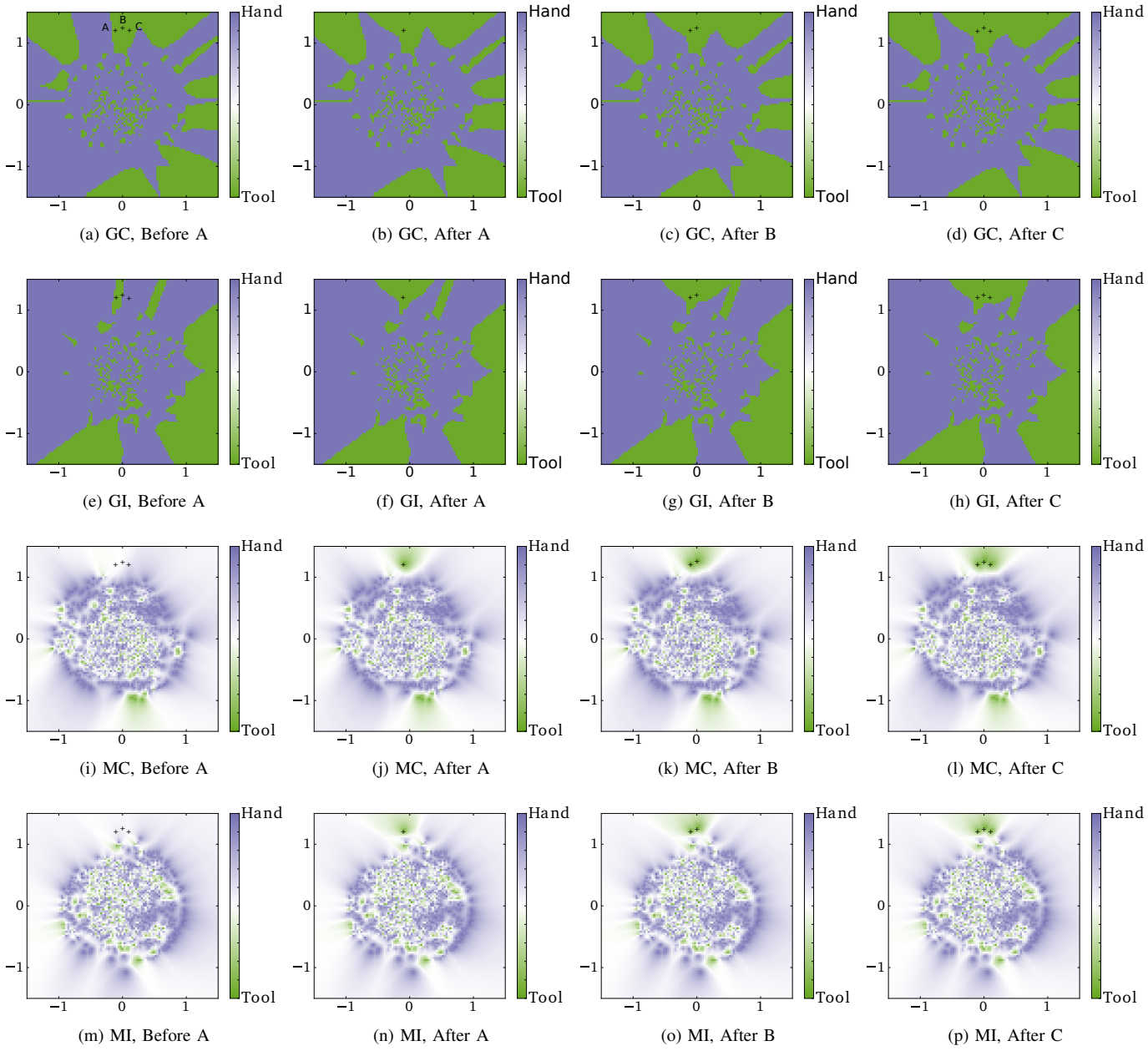
Fig. 4. Evolution of the interest of spaces for one agent of condition MI during 10000 iterations of phase 1.

B and after problem C. The preference is computed as the probability of choosing the hand strategy, and is reported on a two color scale. A completely blue region means that if the block is located in that region, then the corresponding agent would certainly (with probability 1) choose the hand strategy. This is almost the case in conditions GC and GI where the choice is ϵ-greedy with $\epsilon = 5\%$. Similarly, in green regions of those conditions, the choice is almost always for the tool strategy. However, a whiter region (in conditions MC and MI) means that the choice is more balanced, and in completely white regions the choice is equiprobable. It should be noted that the arm is located at position $(0, 0)$, has length 1, and can catch the block within 0.2 so it could theoretically reach the block within a circle of radius 1.2. However, in the 3 problems of phase 2, the block is unreachable directly with the hand. In those problems, the block is located at positions $(-0.1, 1.2)$, $(0, 1.25)$ and $(0.1, 1.2)$ (black dots).

In all conditions (from top to bottom) we can see modifications of the preference around those points across exposure to problems (from left to right), from a hand (blue) to a tool (green) preference. For instance, in condition GC (first row), before phase 2 (first column), this agent already preferred the tool. This is indeed possible because even if during phase 1 we reset the position of the block every 20 iterations to a random position reachable by the hand, this agent could have the time to move the block out-of-reach for the hand and then learn that it could catch it with the tool at that position. This is also part of the reason why success rate increases with experience in all conditions for problem A. Then, after the success to retrieve the toy in problem A (second column), the preference around problem A has changed in a small region around A, but towards the completely different choice: almost always choosing the tool strategy instead of always choosing the hand strategy. The results for the agent in condition GI are similar. However, the results for the agents of conditions MC and MI are different. In condition MC, the agent has no preference in problem A before phase 2, which means that for the first trial to retrieve the toy in problem A, it will choose the strategy randomly, and then the preference might change as the competence value depends on how far from the toy the strategy allowed to reach for. After problem A (second column), the preference changed in a large region around problem A, but this time the change is more gradual, with a high probability to choose the tool strategy only very close to A. The results for the agent in condition MI is similar, but here the preference before phase 2 was for the hand strategy (slightly: 60%, but for other agents it could have been for the tool strategy).

IV. Discussion

We designed an experimental setup where an agent controling a 2D robotic arm could learn two strategies to grab a toy depending on its position: grabbing the toy directly with the hand or first grab a tool to reach for the toy. We defined two dimensions of strategy choice: the type of decision, with a matching law or a greedy choice, and the measure on which to make this choice: the performance of the strategies to retrieve the toy in its current position, or the progress made with each strategy to get the toy. The decision based on the performance measure means that the learner is interested to get the toy, and the decision based on the learning progress means that the toy raises the curiosity of the learner about its affordances or relation with the hand and the stick. The agents have unified learning mechanisms for both free play exploration/learning of tool use from scratch (phase 1) and for exposure to evaluation in lab sessions with an incentive to solve the task (phase 2).



(a) GC, Before A (b) GC, After A (c) GC, After B (d) GC, After C

(e) GI, Before A (f) GI, After A (g) GI, After B (h) GI, After C

(i) MC, Before A (j) MC, After A (k) MC, After B (l) MC, After C

(m) MI, Before A (n) MI, After A (o) MI, After B (p) MI, After C

Fig. 5. Strategy preference maps. Each row corresponds to one condition, plotting the preferences of an agent that succeeded to catch the block on the three problems after having experimented 10000 exploration iterations. Each column shows the map at different times of phase 2: at the beginning of phase 2 before problem A, after problem A, after problem B and after problem C. In those problems, the block is located at positions $(-0.1, 1.2)$, $(0, 1.25)$ and $(0.1, 1.2)$ (black dots). Each 2D map represent the preference between the hand and tool strategies to reach the block depending on its 2D position. In those 2D maps, the arm is located at position $(0, 0)$, has length 1, and can catch the block within 0.2 so it could theoretically reach the block within a circle of radius 1.2 without the tool. The colors reads as follows: a completely blue region (respectively green) means that if the block is located in that region, then the corresponding agent would certainly (with probability 1) choose the hand strategy (respectively tool strategy), and a whiter color means a more balanced probabilistic choice, a pure white being the equiprobable choice. In all conditions (from top to bottom) we can see changes in the preference around those points (from left to right), from a hand (blue) to a tool (green) preference.

First, the success rates in our setup are compatible with the ones of 1.5- and 2.5-year-olds in the experiment of [5], where the success rates increase with experience and also across the successive problems. In this experiment, the toddlers in the control condition did use the other approaches than the tool strategy on more than half the trials after the first time they used the tool strategy in the lab experiment (84% of the trials for 1.5-year-olds, 48% for 2.5-year-olds). However, in the hint and modeling conditions, where the experimenter moreover suggested to use one of the available tools, or actively showed to the infant how to retrieve the toy with the tool, younger infants used other approaches in around 20% of the trials, and older ones in only 4%. In our setup, at each iteration in phase 1 the agents have the choice to explore one of the three available sensory spaces. In phase 2, to model the incentive to get the toy given by the mother in the lab sessions of [5], the agents could only explore the space of the toy, S_{Block}, but could choose either the hand or tool strategy. However, we did not model the hints given by the experimenter in the hint and modeling conditions. Our setup is thus more similar to the control condition of [5] and we observe that only our two conditions using a matching law, MC and MI display a concurrent use of the tool and hand strategies, with smooth evolution to new sensorimotor experience. The behavior of agents in conditions MC and MI are compatible with the overlapping pattern observed with children in the control condition of [5] where the mother just asked the child to get the toy. The fact that the average use of the tool strategy in the control condition did not increase across problems might be because children did not have enough time to discover by themselves that the tool can help to get the toy, without the hint given by the experimenter.

Siegler [2] suggests that the cognitive variability observed in infants could be essential to learning in childhood, and model it as matching law on the competence of the strategies. Our results suggests that an alternative mechanism that was not proposed in Siegler's model could be at play in [5]: strategy selection mechanisms could be based on a measure of learning progress instead of performance.

More generally, a matching law on performance could waste too much experimental trials on high-performing but not improving strategies, even with a novelty bias (that would expire irrespective of progress). On the contrary, a matching law on the monitored learning progress of each strategy could focus the training on low-performing but improving strategies and avoid wasting trials training high-performing but not improving strategies. Indeed, our results also suggests that condition MI could be more beneficial for learning in our setup than condition MC as success rates are slightly better in condition MI. Also, a currently bad strategy could turn out later to be interesting for other related tasks and thus benefit from training. On the other hand, an emphasis on learning progress might too often lead to the choice of an improving strategy that will turn to be sub-optimal. A particular situation with respect to this exploration/exploitation tradeoff is when social feedback is available. In the hint and modeling conditions of [5], the experimenter respectively suggests to use the target tool or actively shows how to retrieve the toy with the tool. The strategic variability is much lower in those conditions, e.g. the 2.5-year-olds used other strategies than the tool one in only 4% of the trials after the first time they use it. We interpret this decrease in variability as the result of the incentive given by the experimenter, supposed to focus attention towards the target tool and to trigger the tool strategy. In our model, the hint condition could be integrated as social bias to select the tool strategy. Also, the demonstration provided by the experimenter in the modeling condition could be added to the sensorimotor models as examples to reach the toy given the trajectory of the tool and the hand, the agent only having to finding motor parameters to realize the hand trajectory (with sensorimotor model 1).

Finally, several strategy selection mechanisms (e.g. based on competence or interest) could be available across all situations, and children could switch between them or combine them depending on the estimated interest of exploration, the desire to actually get the toy, or social cues as the mother or experimenter incentive in Chen & Siegler's experiment.

REFERENCES

[1] J. Piaget, M. Cook, and W. Norton, *The origins of intelligence in children*. International Universities Press New York, 1952, vol. 8, no. 5.

[2] R. S. Siegler, *Emerging minds: The process of change in children's thinking*. Oxford University Press, 1996.

[3] J. Shrager and R. S. Siegler, "Scads: A model of children's strategy choices and strategy discoveries," *Psychological Science*, vol. 9, no. 5, pp. 405–410, 1998.

[4] P.-Y. Oudeyer and L. Smith, "How evolution may work through curiosity-driven developmental process," *Topics in Cognitive Science*, 2016.

[5] Z. Chen and R. S. Siegler, "Across the great divide: Bridging the gap between understanding of toddlers' and older children's thinking," *Monographs of the Society for Research in Child Development*, vol. 65, no. 2, pp. i–105, 2000.

[6] J. Fagard, L. Rat-Fischer, R. Esseily, E. Somogyi, and K. O'Regan, "What does it take for an infant to learn how to use a tool by observation?" *Frontiers in Psychology*, vol. 7, no. 267, 2016.

[7] S. Forestier and P.-Y. Oudeyer, "Curiosity-driven development of tool use precursors: a computational model," in *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 2016.

[8] F. Guerin, N. Kruger, and D. Kraft, "A survey of the ontogeny of tool use: from sensorimotor experience to planning," *Autonomous Mental Development, IEEE Transactions on*, vol. 5, no. 1, 2013.

[9] C. Kidd and B. Y. Hayden, "The psychology and neuroscience of curiosity," *Neuron*, vol. 88, no. 3, pp. 449–460, 2015.

[10] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in Cognitive Sciences*, vol. 17, no. 11, Nov. 2013.

[11] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: learning attractor models for motor behaviors," *Neural computation*, vol. 25, no. 2, 2013.

[12] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, 2013.

[13] M. Rolf, J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, no. 3, 2010.

[14] C. Moulin-Frier, P. Rouanet, P.-Y. Oudeyer, and others, "Explauto: an open-source Python library to study autonomous exploration in developmental robotics," in *ICDL-Epirob-International Conference on Development and Learning, Epirob*, 2014.